DOCUMENT RESUME

ED 024 753 08 VT 001 396

By- Morrison, Edward J.; Lecznar, William B.

Development and Evaluation of an Experimental Curriculum for the New Quincy (Mass.) Vocational-Technical School. Fifth Quarterly Technical Report, the Roles, Characteristics, and Development Procedures for Measures of Individual Achievement.

American Institutes for Research, Pittsburgh, Pa.
Bureau No- BR-5-0009
Pub Date 30 Jun 66
Contract- OEC-5-85-019
Note- 35p.
EDRS Price MF-$0.25 HC-$1.85
Descriptors- *Curriculum Development, Educational Objectives, *Experimental Curriculum, Junior High Schools, Performance Tests, Task Performance, *Test Construction, *Vocational Education
Identifiers- Massachusetts, Project ABLE, Quincy

Technical activity during the period from April 1 through June 30, 1966, was the continued development of junior high school guidance program materials and correlation of arrangements for program implementation, the completion of course and topic objectives in some curriculum areas, and the beginning development of performance measures for verifying student achievement of instructional objectives. Curriculum is being developed in 16 areas including 11 vocational areas, four academic areas, and a new area termed basic technology. In each area, the objectives are stated in terms of the capabilities to be demonstrated by successful students as a result of prescribed learning experiences. A student qualifies for successively higher level jobs or objectives as he progresses through an individually prescribed learning sequence. Performance tests of the capabilities designated as learning objectives will play important roles in diagnosis, achievement demonstration, occupational certification, retention and generalization, orientation and motivation, evaluation and sequencing of learning units, and evaluation of curriculum effectiveness. Organizational arrangements for developing tests have been made, and consideration of the technicalities of standardization, objectivity, format, and representatives is proceeding. Other reports are available as VT 001 392-001 397, VT 004 848, and ED 013 318. (HC)

VT 01396

EDO 247753

FIFTH QUARTERLY TECHNICAL REPORT

Contract No. OE-5-85-019

DEVELOPMENT AND EVALUATION OF AN EXPERIMENTAL CURRICULUM FOR THE NEW QUINCY (MASS.) VOCATIONAL-TECHNICAL SCHOOL

The Roles, Characteristics, and Development Procedures for Measures of Individual Achievement

30 June 1966

U. S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

VT001396

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

DEVELOPMENT AND EVALUATION OF AN EXPERIMENTAL CURRICULUM FOR THE NEW QUINCY (MASS.) VOCATIONAL-TECHNICAL SCHOOL

The Roles, Characteristics, and Development Procedures for Measures of Individual Achievement

Contract No. OE-5-85-019

Edward J. Morrison
William B. Lecznar

30 June 1966

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

American Institutes for Research
Pittsburgh, Pennsylvania

TABLE OF CONTENTS

	Page
FOREWORD	ii
PROJECT OVERVIEW	iii
REPORT SUMMARY	iv
THE ROLES, CHARACTERISTICS, AND DEVELOPMENT PROCEDURES FOR MEASURES OF INDIVIDUAL ACHIEVEMENT	1
Curriculum Structure and Instructional Methods	2
Roles for Performance Measurement	3
Diagnosis	
Achievement demonstration	
Occupational certification	
Retention and generalization	
Orientation and motivation	
Evaluation and sequencing of learning units	
Evaluation of curriculum effectiveness	
Summary	
Characteristics of the Measures	7
Types of decisions facilitated	
General and procedural characteristics	
Validity	
Reliability	
Scoring	
Summary	
Development Procedures	14
Organizational arrangements	
Standardization, objectivity, formats	
Representativeness	
REFERENCES	19
PLANS FOR NEXT QUARTER	21
APPENDIX A. PROFICIENCY TEST DEVELOPMENT	
APPENDIX B. DESCRIPTION OF TYPES OF QUESTIONS	

FOREWORD

This report, submitted in compliance with Article 3 of the contract, reports on technical activities of Project ABLE during its fifth quarter of operation, 1 April through 30 June 1966. A brief overview of the project is presented first, followed by a report summary. The major portion of the report is a discussion of the development of performance measures to be used to assess students' achievement of the objectives of instruction. Project plans for next quarter are outlined.

OVERVIEW: Project ABLE

A Joint Research Project of: Public Schools of Quincy, Massachusetts
and American Institutes for Research

Title: DEVELOPMENT AND EVALUATION OF AN EXPERIMENTAL CURRICULUM FOR THE NEW QUINCY (MASS.) VOCATIONAL-TECHNICAL SCHOOL

Objectives: The principal goal of the project is to demonstrate increased effectiveness of instruction whose content is explicitly derived from analysis of desired behavior after graduation, and which, in addition, attempts to apply newly developed educational technology to the design, conduct, and evaluation of vocational education. Included in this new technology are methods of defining educational objectives, deriving topical content for courses, preparation of students in prerequisite knowledges and attitudes, individualizing instruction, measuring student achievement, and establishing a system for evaluating program results in terms of outcomes following graduation.

Procedure: The procedure begins with the collection of vocational information for representative jobs in eleven different vocational areas. Analysis will then be made of the performances required for job execution, resulting in descriptions of essential classes of performance which need to be learned. On the basis of this information, a panel of educational and vocational scholars will develop recommended objectives for a vocational curriculum which incorporates the goals of (a) vocational competence; (b) responsible citizenship; and (c) individual self-fulfillment. A curriculum then will be designed in topic form to provide for comprehensiveness, and also for flexibility of coverage, for each of the vocational areas. Guidance programs and prerequisite instruction to prepare junior high students also will be designed. Selection of instructional materials, methods, and aids, and design of materials, when required, will also be undertaken. An important step will be the development of performance measures tied to the objectives of instruction. Methods of instruction will be devised to make possible individualized student progression and selection of alternative programs, and teacher-training materials will be developed to accomplish inservice teacher education of Quincy School Personnel. A plan will be developed for conducting program evaluation not only in terms of end-of-year examinations, but also in terms of continuing follow-up of outcomes after graduation.

Time Schedule: Begin 1 April 1965
Complete 31 March 1970
Present Contract to 30 June 1967

REPORT SUMMARY

During the present reporting period, technical activity was directed primarily to (1) continued development of junior high guidance program materials and completion of arrangements for program implementation, (2) completion of course and topic objectives in some curriculum areas, and (3) the beginning of development of measures for verifying students' achievement of instructional objectives. The present report is concerned with achievement measures. It reviews the curriculum structure and instructional methods which have been planned and identifies a number of important roles for which achievement measures are needed. The technical requirements for measures employed in those roles are examined and the procedures for developing such measures are discussed.

During the next quarter, test development will occupy a greater proportion of total activity. Selection and development of instructional materials, aids, and procedures will continue concurrent with the development of measures. Junior high guidance preparations will be completed and the program will be initiated.

THE ROLES, CHARACTERISTICS, AND DEVELOPMENT PROCEDURES FOR MEASURES OF INDIVIDUAL ACHIEVEMENT

Perhaps the most distinctive characteristic of Project ABLE to date is its persistent focus on the performance capabilities of students. This emphasis was established at the outset by the statement of the project's purpose which was, in part, to evaluate the effectiveness of a curriculum derived explicitly from the behavior desired of graduates. It was taken as fundamental that education aims primarily to produce learning by students; that learning involves changes in the capabilities of students, that is, that a student has learned when he can demonstrate a capability which he could not demonstrate before the learning experience; and that the basic design task of the project was to select the demonstrable capabilities desired of students and to establish conditions under which those capabilities could be acquired efficiently.

Adherence to this primary purpose, and to the rather simple assumptions associated with it, has led us over new routes to results quite different from the usual products of curriculum development. Previous reports (American Institutes for Research, 1965a, 1965b, 1965c, 1966) describing the development procedures, instructional objectives, curriculum outlines, and guidance programs of the project reveal the differences in curriculum design. That work can not be recounted here, but it should be noted that the statement of instructional objectives in behavioral terms was the key to most differences between Project ABLE products and those more commonly obtained. Vague and uncertain statements about what the student should learn were avoided in favor of clear statements about what he should be able to do. Objectives were identified as the content of the curriculum, and content was distinguished thereby from the conditions under which learning would take place (e.g., teachers' activities, instructional methods, materials, aids, procedures).

The design of the curriculum, then, has proceeded in accordance with the original purpose. But the implementation of the curriculum and the

evaluation of its effectiveness require means for assessing the performance capabilities acquired by the students. The remainder of this report is devoted to consideration of the problems of performance measurement. Following a brief description of the curriculum structure and the educational methods which are relevant to the problems of measurement, the discussion is organized around three major topics. First to be considered are the roles assigned to performance measures. This review of functional requirements leads to an examination of the principal technical characteristics which the measures must have to play their intended roles. Finally, specific procedures are summarized for developing operational measures.

Curriculum Structure and Instructional Methods

Curriculum is being developed in 16 areas: 11 vocational areas, 4 "academic" areas, and a new area called basic technology. In each, the content will consist of a set of objectives stated in terms of the capabilities to be demonstrated by successful students as a result of prescribed learning experiences. The objectives are being organized hierarchically. That is, each area has a set of "course" objectives at the top of the hierarchy. These are the end capabilities toward which all earlier learning is to be directed. Each course objective has subordinate "topic" objectives which are statements of prerequisite capabilities. Objectives subordinate to topic objectives are provided when required. The learning sequence will extend at one end to the lowest capability level expected of entering students, and at the other end to the highest capability level for which training is to be provided. The curriculum structure conforms in general with the hierarchical concepts described by Gagné (1965).

The sequence of learning objectives is being defined in accordance with two major considerations. The first consideration has been suggested above in reference to prerequisite capabilities. That is, some capabilities occur later in a sequence because the student cannot acquire them unless he is able already to do the things specified in earlier objectives. For example, the student cannot learn to solve systems of linear equations until he has learned to perform simple algebraic manipulations. In the vocational areas, a second reason for the sequences is that the learning objectives

are intended to parallel the structure of jobs selected for training. Thus, in each vocational area, a sequence of jobs was chosen such that a large proportion of the skills and knowledges of any job also are required for successful performance of jobs later in the sequence. In this arrangement, a student qualifies for successively higher-level jobs as he progresses through the learning sequence. Within broad limits, each student thus would have marketable skill whenever he leaves school.

The curriculum structure is intended to be at least compatible with individualized instructional methods. It is planned that each student will proceed through an individually prescribed learning sequence, advancing to the next objective as he demonstrates that he has acquired the prerequisite capabilities. It is planned also that lectures by the teacher will be minimized in favor of individual study, small group discussions, demonstrations, and tutorial work.

With this brief review of curriculum structure and instructional methods as background, we turn now to consideration of the roles of performance measurement.

Roles for Performance Measurement

This section is concerned with the several roles which it is expected that performance measures will play in the conduct and evaluation of the experimental curricula. These roles are the "why" of performance measurement, and the descriptions which follow indicate the uses to which we expect to put the results of measurement.

Diagnosis. If students are to work their various ways through individually prescribed sequences of hierarchical objectives, it is important that the first event in each student's experience with a curriculum area be a determination of his present capabilities. What is required is a diagnostic report which locates the student's proper starting place in the curriculum by identifying the relevant capabilities he can demonstrate and those he has not yet learned. With this information available, the teacher can identify the learning assignments which the student should attempt in order to proceed efficiently toward his educational objectives. This

diagnostic test may reveal that the student lacks some essential capability normally acquired in junior high school. In such a case, the appropriate assignment would provide the student with the opportunity to acquire that capability before attempting to meet objectives which depend on the skill or knowledge which is missing from his repertoire. In other cases, entering students may demonstrate some capabilities which are well in advance of the usual starting place in the curriculum. The appropriate assignment for these students would provide them with the opportunity to build on their past learning without having to go through material and exercises which would not add to their existing capabilities. The diagnostic measurement of performance capabilities is an important role because it provides the basis for individually prescribed sequences of learning assignments.

Achievement demonstration. A second role for performance measures is closely allied with the first. They are expected to function as the means whereby the student demonstrates achievement of each learning objective. In the instructional procedures being planned by the project, each learning assignment to a student would include a statement of the end performance to be demonstrated, the important conditions under which the demonstration is to take place, and the criteria by which the performance is to be judged. The student would take the test on a learning unit when he believed himself able to pass. If he succeeded, he would progress to another learning task. If he did not pass, he would return to the same assignment, or to remedial or alternative assignments as necessary, until he could demonstrate that he had accomplished the assigned learning. The performance measures thus would function as the means by which students demonstrate at each step their readiness to progress in the curriculum.

Occupational certification. Performance measures are expected to play an additional role in vocational areas of the curriculum. Thus, it is planned that as a student passes each test, he provides evidence thereby that he has a capability required for competence in one or more occupations. When he has demonstrated all of the required capabilities, he is eligible for certification by the school as competent in an occupation. The vocational performance measures thus would be the means by which students

earned official recognition of their marketable capabilities. A student might qualify for several certificates in the course of his secondary education. Normally, he would be awarded only the last (or highest level) one earned, though any earned certificate could be supplied on the basis of his record.

Retention and generalization. Each of the roles described thus far is a measurement primarily of capability deliberately acquired, and is taken at the point of first mastery. That is, the student works on acquisition of a capability and then promptly demonstrates his mastery of it on a test designed for that purpose. It is reasonable, however, to measure two additional aspects of a student's capabilities at selected points in his development. Such a point might be at the completion of a number of assignments which are coordinate objectives, all prerequisite to a major learning task. Since these prerequisite learnings would be accomplished over a period of some time, it might be important to verify the student's retention of these previously demonstrated capabilities before he went on. Any important deficiency then could be remedied before the student attempted the major learning task for which his area of deficiency was a prerequisite.

In addition to verifying the retention of previous learning, different performance measures could be introduced at selected points to assess the student's ability in areas not covered directly by his previous assignments. Such tests of the generalization of learning would be useful in deciding whether a student would profit more from assignments designed to broaden his capabilities in some portion of a course of study or more from proceeding to advanced levels of the sequence. These measures also would provide information about the extent to which the curriculum contributed to the achievement of educational outcomes other than those specified by the objectives. This matter is related to a later discussion of the role of performance measures in curriculum evaluation.

Orientation and motivation. It is expected that an objective and its passing requirement stated in advance for the student in performance terms would provide him with an unusually clear goal which may be attained in a modest amount of time. Since the relation of each individual objective to the student's longer range goals would be demonstrable through a sequence

of learning objectives, the necessity and relevance of each achievement should be clear. Further, the outcome of each test of the student's learning would be clear to the student and to the teacher. The combination of clearly-defined, relevant requirements, unambiguous evaluation, and frequent opportunity to achieve is expected to enhance the student's motivation for learning.

Evaluation and sequencing of learning units. In the curriculum development process, every effort is being made, of course, to devise effective learning units and to arrange them in hierarchical sequence. However, it is easy to err in this process. Ineffective units can appear and units can be arranged in erroneous sequences when the development depends on rational analysis only. The performance data collected during tryout of the curriculum are expected to provide an empirical basis for evaluation of the effectiveness and sequencing of the units. Such findings as unexpectedly long times to complete units, repeated failure to pass the unit tests, and large proportions of failed first attempts all would indicate defective units. The sequence in which units are arranged also can be evaluated from the results of unit performance tests (Gagné, 1966). The result expected from a proper sequence is that all, or nearly all, of the students passing a unit also would pass units presumed to be its prerequisites. Pass-fail data arranged in a student-by-unit matrix and data on proportions of students passing each unit provide evidence as to the tenability of the initial sequence, possible rearrangements, and the need for additional units. Performance measures therefore are expected to play an important role during the tryout period by facilitating the evaluation and revision of the experimental units and their sequential arrangement.

Evaluation of curriculum effectiveness. Clearly, the performances of students on measures of their capability for tasks defined as learning objectives are basic data inputs to curriculum evaluation. They provide an answer to the fundamental question, "Did students learn what we intended that they learn?" But many other questions must be asked in evaluating the effectiveness of the curriculum, and some of these questions will require that other data be gathered on students' capabilities. A previous section

identified the need for performance measures designed to assess the extent to which the curriculum provides extra values through generalization of learning and through acquisition of "incidental" skills and knowledges. It is planned that such measures will be used and that they will contribute to the evaluation of the extent to which many important educational objectives, not stated as specific objectives for the curriculum, are met (Cronbach, 1964).

Summary. Performance tests of the capabilities designated as learning objectives for the student are expected to play important roles in:

- diagnosing the initial learning status of each student and prescribing individually appropriate sequences of assignments.
- demonstrating that unit objectives have been met and that the student is ready to proceed to another unit.
- certifying students in occupations.
- verifying retention of previously demonstrated capabilities.
- orienting students to and motivating them for learning.
- evaluating individual learning units and their sequencing.
- assessing curriculum effectiveness.

Other performance measures are expected to be used in assessing the generalization of learning, the acquisition of "incidental" skills and knowledges, and the extent to which other important educational outcomes are achieved.

Characteristics of the Measures

The characteristics needed in a measurement depend upon the uses for which the measure is intended and upon the operational conditions under which the measurement will be taken and used. The preceding sections have described both the uses and the operational conditions planned by Project ABLE. This section will consider the implications of those conditions and uses for the kinds of measures we need.

Types of decisions facilitated. Cronbach (1960, 1964) has distinguished between tests in education according to the kinds of decisions they are expected to facilitate. Thus, tests are used to make selection and classification decisions about individuals, to evaluate and revise curricula, to make decisions for administrative regulation, and to test scientific hypotheses. Glaser and Klaus (1962) present a similar analysis and also describe quality control and system evaluation functions, which can be considered mixtures of Cronbach's decision types. In Project ABLE, the largest number of decisions by far will concern individual students. Tests will be used in decisions as to whether a student is ready to enter a sequence of study, which of the available assignments he should attempt, whether he has met the objective of a particular assignment and should be given additional work, or whether he needs to repeat an assignment. The results from these tests also will be major input data during the tryout period for decisions about the curriculum, including the revision and sequencing of learning units. Later, they will assist in evaluating the effectiveness of the curriculum.

Tests used for decisions about individuals differ from other tests in two major respects. First, decisions about curriculum or about administrative matters usually can be based upon means of test data from samples of students. Not all students must be tested and unsystematic errors made in measuring the capabilities of individual students need not affect the appropriateness of decisions, since these errors tend to offset each other in the averaging process. When decisions are to be made about individuals, however, errors in measuring the capabilities of those individuals must be minimized. Secondly, it is more important in individual decision situations that the assessment be recognizable by the student as a fair and adequate measure of capabilities relevant to his educational goals. These two characteristics, individual accuracy and recognizable adequacy, will be important considerations in the following pages.

Since the majority of tests must be devised to facilitate decisions about individuals, and since the data from these tests will serve secondarily as a major part of the basis for decisions on curriculum revision and administrative regulation, we will direct our attention in the remainder of this report to the characteristics needed for such tests. Insofar as additional

proficiency measures are required for curriculum evaluation purposes, they are best selected or devised and discussed as part of an integrated evaluation program which will be the subject of a later report in this series. It might be noted that it would be inappropriate to develop the number and kinds of tests needed for individual decision if only curriculum and administrative decisions were required. However, since these more demanding tests must be developed by the project, no inefficiency is incurred and their use for purposes other than individual decisions involves no technical hazard.

General and procedural characteristics. The curriculum structure and instructional methods described earlier require that a very large number and variety of tests be devised. Achievement of any course objective is expected to require prior acquisition of numerous and diverse capabilities. Even though an end-of-course objective might be met by demonstrating the ability to produce some kind of machined part, for example, the constituent capabilities which must be acquired first may well call upon a wide range of psychological processes, response patterns, and stimulus contexts. Appropriate tests of the student's capabilities during the learning process must be equally diverse. Our tests must assess the capability for which training was devised, using whatever supporting materials and conditions are appropriate. We would expect to use paper-and-pencil, equipment, job samples, oral reports, simulation or whatever is essential, being guided in our choice by the stimulus context, the psychological processes, and the response modes demanded by the performance objective.

Not only will the tests exist in great diversity, but they will be administered and interpreted by many different teachers. Further, results from the tests will be collected and analyzed by a separate research staff concerned with decisions other than the assessment of individual students. These requirements make it clear that procedures for administering and scoring each test should be standardized and that the test results should depend only minimally on the observer. Through standardization and objectivity we may hope to succeed in orienting and motivating the students, in providing fair and unambiguous results acceptable to student and teacher, and in providing reports on learning achievement which are sufficiently dependable for our operational and research purposes.

Validity. The point has been made several times in this paper that it is important that tests of individual capability measure performances which are relevant to the students' educational objectives. This is the essential question of test validity. A test is valid to the extent that it does the job it was intended to do, in this case to report on each student's achievement of the stated objectives.

Long-term objectives for the student, stated at the outset of the project, included vocational competence, responsible citizenship, and self-fulfillment. These objectives were reasonable and useful as goals, but they were not satisfactory as working objectives for curriculum development or as criteria for achievement test validation. First, of course, though our "real" interest might be in the student's performance after graduation, as indicated by these goals, we could not wait several years for students to demonstrate their accomplishments. Secondly, the goals as stated lacked specification in terms of the performance capabilities they are intended to imply. As Cureton (1951, p. 641) points out, such goals are merely labels representing abstract concepts and summarizing the behavior of persons whose actions within some defined series are characteristically successful. Objectives were needed which were closer in time to student learning and which specified the actions or performances which define the concepts of vocational competence, responsible citizenship, and self-fulfillment. It was apparent that our ultimate goals could not be used directly as criteria against which to validate our curriculum or our achievement tests.

Using a procedure described elsewhere (American Institutes for Research, 1965b), specific objectives were derived from the general goals by logical procedures. These more proximate, intermediate objectives describe the capabilities to be acquired by students in units and in courses of learning. They constitute a definition of the capabilities which our analysis indicates are essential to achievement of the long-range goals and which are feasible objectives within the public school context. The statements of learning objectives include a description of the performance, the criteria for judging success, and the important conditions under which the performance is to take place. The objectives are intended to imply directly how achievement should be measured. Thus, the topic and course objectives are the criteria for

evaluation of test validity. The relevant achievement test for any objective thus is performance of examples of the criterion task as described in the learning objective. The empirical relation between achievement of curriculum objectives and achievement of the long-range goals, that is, the validity of the achievement tests for predicting success in later life, must be dealt with in long-term, follow-up studies. It is not considered further in this report.

Since the test performance is intended to be a representative performance of the criterion task, the question of test validity becomes one of the representativeness of the test tasks. Thus, if the student's objective were to be capable of solving sets of simultaneous linear equations, then particular sets of equations would be needed for the test performance. The test would be considered valid only if the test equations fairly represented the universe of equations described by the objective. In addition, the test should be representative of the criterion with respect to important conditions of the task. In the example cited, the time allowed for solutions, the accuracy required in answers, the amount and portions of the solution which must be displayed, etc., are possible criterion conditions which should be fairly represented in the test.

The question of the representativeness of test tasks has practical, methodological, and theoretical aspects. Thus, criteria can be imagined for which a fully representative test would be virtually endless because the examples of the criterion task, or the conditions under which the task would be performed, are extremely diverse. On the other hand, criteria can be written which are so specific as to include only one example or test task. As a practical matter, neither criterion serves well as an educational objective or as a criterion for measurement. Each would be modified to encompass more appropriate amounts of learning and testing. Still, these extreme types of criteria raise the theoretical problem of defining representativeness and the methodological problem of devising methods for selecting a set of tasks demonstrably representative of the criterion. These same problems of representativeness or task identity appear in contexts other than achievement testing, notably in the design of training

devices (Gagné, 1954) in the various applications of system simulation (e.g., Davis & Behan, 1962), and in the analysis of jobs for core content (Altman & Gagné, 1965). No formal and generally applicable theory or method is available for assuring that test tasks are truly representative of the criterion, though Altman (1966) describes a psychological-process x content model with interesting possibilities. This does not reduce the importance of representing the criterion faithfully in our tests of individual achievement, nor the need for deliberate practical attempts to assure that the achievement tests are relevant to the criteria.

Reliability. Every measurement errs to some degree in estimating the true value of the variable measured. Repeated sets of measures of the same individuals never exactly duplicate one another. Every set of measurements thus is unreliable to some degree. The degree of unreliability in a measure is of practical importance because it determines the confidence with which the measure may be used as a basis for decisions. If a measure is sufficiently unreliable, it is worthless as a basis for decision, no matter how relevant (valid) the test tasks may be with respect to the criterion.

In an earlier section, it was pointed out that tests used in decisions about individuals should evaluate individual performances with less error than would be tolerable were the same tests to be used only for curriculum and administrative decisions. But no precise statement has been or can be made at this time as to the minimum level of reliability which must be achieved by the measures in this project. Such statements can be developed from specifications of the size of test score differences which must be detected and of the risks of error which can be tolerated (e.g., Kelley, 1927). However, in a practical situation, such as an operating school, preset standards for discriminations and risks may be of little value. Actions must be taken on the best available basis, even if the risks are larger than desired. In the present project, it seems clear that highly reliable measures of individual achievement should be a goal for test development. Such measures would contribute substantially to the efficiency of the curriculum operation and to the fairness with which each student is is dealt. The basic objective should be for students only rarely to repeat

or pass a learning unit as the result of testing error. But the goal of high reliability should not be achieved through significant sacrifice of relevance in the measures nor through important restriction of the learning activities. Fortunately, the curriculum structure and the procedures planned in individualizing instruction can tolerate some unreliability in the achievement tests. Long-term difficulty for a student is unlikely to result from an occasional error in evaluating his performance on individual learning units which are relatively short. An erroneous "failure" decision can be overcome as soon as the student decides he is ready for retest. An erroneous "pass" decision should result in detectable difficulty with the next higher learning unit and precipitate correction of the assignment error.

Scoring. Objectives for learning units, which serve as criteria, include specifications for a satisfactory performance. The test of a student's achievement of the objective is required in this program to produce only a pass-fail score based on whether his performance meets or exceeds the specified standard. For purposes of validity and reliability in measurement, several example performances may be required of the student in test, so that considerable data should be available to support the pass-fail decision and other analyses. But the primary test score need be only dichotomous. The purpose of each test is to compare a student's performance with an *a priori* standard, not to compare his performance with that of other students or with established norms. The measures required in this program are an example of "criterion-referenced measures" (Glaser & Klaus, 1962) which indicate the content of the student's behavioral repertory without reference to the performance of other persons.

The pass-fail score requires that students be sorted into only two groups. Were we to require a finer sorting (say, into low fail, fail, pass, high pass), a larger number of sorting errors would be expected. If we required a sort into N groups, where N is the number of students, the resulting rank ordering of students would be expected to include yet a larger number of errors. The pass-fail score is expected, therefore, to produce the minimum error and the highest reliability of the scores which could be used. While this is not the major reason for its use, it is a welcome result.

Summary. Plans for the structure of the curriculum for the instructional procedures, and for the roles to be played by performance measures require that many diverse tests be devised which are:

- adequate to support educational decisions about individual students.
- reasonably standardized with respect to administration, scoring, and interpretation.
- representative of the universe of tasks defined by the objectives for learning units.
- as reliable as practical constraints permit without significant sacrifice of validity.
- scored by reference to the criteria provided by the learning unit objectives.

Development Procedures

The discussion of tests so far has considered the educational arrangements within which tests will be used, the roles they will be expected to play, and the technical characteristics they consequently must display. This section considers briefly some major aspects of the procedures being employed in test development.

It should be noted that only a few of the curriculum areas have reached the point of developing proficiency measures as of this reporting date. Relatively small amounts of test material could be displayed and our technical and operating procedures still are in the shakedown process. However, the general outline of our modes of operation and our handling of central problems can be described.

Organizational arrangements. The professional staff of the project includes 16 members of the faculty of the faculty of the Quincy Public Schools and three research people from A.I.R. Each faculty member has responsibility for curriculum development in one area and is, by training and occupation, a specialist in his area. Faculty members provide the

project with subject-matter knowledge, technical skills, and teaching experience. Each faculty person in a vocational area also has work experience in his area of responsibility. Many Quincy faculty members not assigned to the project are nonetheless available to the project for consultation, review of products and specific assignments for which they are especially qualified. Experienced A.I.R. people provide the project with skills and knowledge in methods of research and in educational and psychological measurement. Each task in the program is attacked as a cooperative effort of these two groups.

In the development of proficiency measures, research members are responsible for analysis and definition of the technical requirements for the measures, for devising or selecting the procedures to be used in developing the measures, for preparing procedural and technique guides, and for providing test writers with direction, assistance and technical review. In working with faculty, research people are especially concerned with the behavioral aspects of the measures. That is, they attend to the problem of assuring that the psychological processes, response modes, and stimulus contexts required in the criterion performance are represented appropriately in the test task. Faculty specialists are responsible for developing the test items in accordance with requirements. In this work, they are especially concerned with knowledge and skill content of the tests.

Standardization, objectivity, formats. The procedures and techniques employed in preparing the various kinds of test items are standard in test development practice (cf., Adkins, 1947; Lindquist, 1951). Project research members have prepared abbreviated guides (examples are shown in Appendices A and B) for use by the faculty specialists and have augmented these with instruction, consultation and review of finished items. Objective scoring is intended for all items, though some complex performances will be evaluated by use of checklists and, in rare cases, by rating methods. Appropriate test materials will be supplied to teachers with each set of learning unit materials and teachers will be instructed in their use so as to enhance standardization of testing procedures, scoring, and interpretation.

Representativeness. The most difficult development task is to assure that test tasks do in fact provide a representative demonstration of the capabilities defined as objectives for learning. As mentioned in an earlier section, no formal method or theory is available whereby test representativeness is guaranteed. Consequently, we must depend basically upon the combined judgments of research and faculty people to produce valid measures. Though risk is involved in this logical process, the nature of the objectives and the systematic use of a partial frame of reference help to objectify the procedure. The problem can be described in the following two parts.

1. Assuring that the test task is an example of the criterion performance.

The appropriate test task is quite apparent in many instances. For example, achievement of an objective which states, with appropriate additional specification, that the student should be able to measure voltage using a given meter is assessed by requiring the student to do exactly that. Similarly, appropriate examples are written rather easily for such common objectives as solving equations, listing causes, punctuating, or reciting physical laws.

In other instances, test tasks are less easily certified as examples of the criterion. Usually, this is because the behavioral statement in the objective is not sufficiently specific. Consider a fictitious objective which requires that the student know the proper nomenclature for each part of machine X. Should the test require the student to write from memory a list of the parts? Mark the names of parts in a longer list? Say the correct name when the instructor points to the part? Write the names of parts indicated in a picture of the machine? Several of these? Does it matter which is used? It is clear that these possible test situations involve different psychological processes, different response modes, and different stimulus presentations, though all are directed at the same "knowledge" (nomenclature of the parts). Unless the statement of criterion behavior provides the relevant information, there is no basis for choice among the possible item types or for asserting that any of the items is the task intended by the objective.

In practice, the appearance of such an item-objective pair would result in review and restatement of the objective so that the appropriate examples could be identified. But the essential requirement for assuring that test situations are examples of the criterion is a standard frame of reference defining the dimensions of important variation among performances. It is helpful to know that one should compare the test and criterion performances with respect to their response modes, psychological processes, and stimulus contexts, but what is a useful way to distinguish between, say, response modes? When are two modes functionally equivalent? When does success in one mode imply capability in the other? In the absence of a comprehensive basis for such decision, we must rely upon persons experienced in the analysis of behavior to render judgments with respect to specific items. Considerable assistance is provided in part of this task by reference to the hierarchical categories of psychological processes proposed by Altman (1966), who also has described the relations between these processes, the learning categories of Gagné (1965) and others (Melton, 1964), and classes of behavioral error.

In passing it should be noted that one type of error frequently detected in item review (and not only in this project) is to specify a test item which requires a verbalization about the desired performance rather than the performance itself. Thus, instead of requiring the student to make a machine set-up as required by the objective, he might be asked to list the steps in that procedure. Frequently, especially in "academic" areas, verbalizations are perfectly appropriate objectives and are properly tested for by asking for verbal performance. In some cases, verbalization about or simulations of non-verbal performance are the only reasonable ways to estimate achievement of an object. Proper behavior under certain emergency conditions would be an example of such instances. But there are many instances in which a verbal description of the performance is by no means equivalent to the performance itself and our tests are intended to exclude errors of this sort.

2. Assuring that the test tasks adequately represent the universe of examples.

Assuming that a properly stated objective defines a capability for a class of performances, the test should include demonstration of capability

across the important varieties of the performance. For example, if one data processing criterion capability were to perform all standard card sorting operations with a particular class of machines, we would want to be sure that the student could sort alphabetically as well as numerically, could produce the major types of card sorts, could handle large as well as small numbers of cards, etc. This is the problem of representing in a few test items all of the important specific performances of which the student should be capable if he has met the learning objective.

It is not necessary or desirable in all instances to include routinely every minor variation in tasks. Often, testing near the extremes of a dimension of variation or including all important dimensions in one task is more efficient and quite adequate. The important consideration is whether, having succeeded at the test tasks, the student has demonstrated his capability for performing all important instances of the criterion capability.

We are dependent largely upon the knowledge of faculty specialists and their colleagues in identifying the important dimensions of variation among tasks. Item review by the research staff and discussion with the specialists provide a check on the adequacy of task sampling. The hazard is that some tasks may be omitted in favor of more easily tested items or tasks which are more familiar to the individual preparing the test. The lack of formal methods and theory to support the comparison of human performances is again a handicap and is dealt with here in a manner analagous to the procedure described in the first part of this section on representativeness.

REFERENCES

Adkins, Dorothy C. Construction and analysis of achievement tests. Washington: U. S. Government Printing Office, 1947.

Altman, J. W. Research on general vocational capabilities (skills and knowledges). Pittsburgh: American Institutes for Research, March 1966.

Altman, J. W., & Gagné, R. M. Research on general vocational skills. Pittsburgh: American Institutes for Research, October 1964.

American Institutes for Research. Project ABLE: First quarterly technical report. Pittsburgh: Institute for Performance Technology, June 1965. (a)

American Institutes for Research. Project ABLE: Second quarterly technical report. Pittsburgh: Institute for Performance Technology, September 1965. (b)

American Institutes for Research. Project ABLE: Third quarterly technical report. Pittsburgh: Institute for Performance Technology, December 1965. (c)

American Institutes for Research. Project ABLE: Fourth quarterly technical report. Pittsburgh: Institute for Performance Technology, March 1966.

Cronbach, L. J. Essentials of psychological testing. New York: Harper & Row, 1960.

Cronbach, L. J. Evaluation for course improvement. In R. W. Heath (Ed.), New curricula. New York: Harper & Row, 1964.

Cureton, E. E. Validity. In E. F. Lindquist (Ed.), Educational measurement. Washington: American Council on Education, 1951.

Davis, R. H., & Behan, R. A. Evaluating system performance in simulated environments. In R. M. Gagné (Ed.), Psychological principles in system development. New York: Holt, Rinehart & Winston, 1962.

Gagné, R. M. Training devices and simulators: Some research issues. American Psychologist, 1954, 9, 95-107.

Gagné, R. M. The conditions of learning. New York: Holt, Rinehart & Winston, 1965.

Gagné, R. M. Curriculum research and the promotion of learning. Invited address to meetings of American Educational Research Association, Chicago, February 1966.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), Psychological principles in system development. New York: Holt, Rinehart & Winston, 1962.

Kelley, T. L. Interpretation of educational measurements. Yonkers, N. Y.: World Book, 1927.

Lindquist, E. F. (Ed.) Educational measurement. Washington: American Council on Education, 1951.

Melton, A. W. (Ed.) Categories of human learning. New York: Academic Press, 1964.

PLANS FOR NEXT QUARTER

The following activities are planned for the quarter ending 30 September 1966:

1. Development of performance measures will continue, becoming the primary activity in most curriculum areas.
2. Selection and development of instructional materials, methods, aids, and procedures will continue in some areas concurrent with the development of measures.
3. Selection, organization, and development of curriculum topics will continue in "academic" areas in accordance with conclusions reached in faculty meetings with the Advisory Panel during May.
4. Materials, staff training, and implementation arrangements will be completed for the junior high guidance program and the program will be initiated.
5. Development of plans for the senior high guidance program will continue.

APPENDIX A

PROFICIENCY TEST DEVELOPMENT[1]

1. The primary purpose of proficiency measurement is to provide an objective assessment of criterion behavior which, in Project ABLE, is attainment of a series of educational objectives. These goals have been stated as topic objectives relating back to a set of tasks (through course objectives) for specifically selected jobs. Since these topic objectives have been developed to detailed levels of behavior by asking for each performance in turn, 'What kinds of previously learned capabilities need to be assumed if the person is to learn <u>this</u> capability under a single set of learning conditions?", it should now be possible to reflect the criterion as stated (or implied) in each topic objective into one or more items that will measure a student's success or failure in achieving the specified capability or educational goal.

2. Within the general purpose of assessing criterion behavior, we have two associated reasons for measurement. The first is to assess present performance for initial <u>placement</u> into that point of the curriculum which the student is capable of completing satisfactorily without retracing or repeating previously acquired skills. The second is to determine <u>performance</u> adequacy of terminal behavior specified as the final training product; the proficiency test in this case may sample situations other than those explicitly covered in training so as to evaluate the extent to which specific behaviors have been generalized to a variety of potential job situations.

3. The ease with which the development of a proficiency measure can be carried out is dependent upon (1) the <u>complexity</u> of the behavior involved,

---

[1] Adapted from Adkins, Dorothy C., <u>Construction and analysis of achievement tests</u>. Washington: U. S. Government Printing Office, 1947.

(2) the explicitness with which the behavior has been defined, and (3) the accessibility of the behavior to observation. Because our effort from the start has been directed toward specific job-oriented capabilities, it may be feasible and desirable to develop proficiency tests directly from topic objectives prior to specification of curriculum content and selection of materials which in themselves will be oriented toward the same capabilities as the proficiency tests.

4. To provide usable information, the proficiency measures must be objective and quantified. They may be in the form of test items (probably multiple choice recognition to simplify the quantification), checklists covering demonstrated performance, or rating scales (the least reliable or desirable). To meet our stated needs, the measures will be criterion referenced (an absolute standard of proficiency to be met by each individual student) rather than norm referenced (each student compared to other students). Thus the items will not be written at varying levels of difficulty. They must be directed, however, to the outcomes specified in the topic objectives, or at least to the desired capabilities if these are not clearly specified in the topic objectives. The proficiency measure will tell both the student and the teacher whether a capability has been achieved, and that knowledge needed to proceed has been acquired; it may also be used later in specifying curriculum content. The test items must be given thoughtful, careful writing to satisfy the requirements for their several later applications.

5. A procedural outline that may help in preparing proficiency measures is provided below.

a. Begin with the lowest skill level job and work successively one job at a time to the highest level. Since jobs were selected to build on the skills of prior levels, this will give an initial sequencing of capabilities.

b. For each topic objective written to each job, note carefully the critical behavior (capability to be established) and prepare two or more items by which student achievement

of each capability can be evaluated. More than one item is required to increase accuracy of measurement (allow for possible "bad" items).

c. Since topic objectives were completed in varying degrees of acceptability with respect to specificity and level of capability, it may be necessary to review content of each test item to see if a previously unidentified (no topic objective written) capability is assumed. (In the attached sample, the first item was written to a specific topic objective on selecting the proper grinding wheel; on inspection another item was seen as necessary to establish a capability with reference to tensile strength of metals.)

d. Write each item on an individual sheet, 5" x 8" in size, to facilitate later sequencing, editing, and typing. Be sure to complete all identifying data. Use the reverse side of the form for continuation of an item if necessary, but only one item to a sheet.

e. Indicate by use of a check or star the correct alternative; use upper case letters to designate the alternatives.

f. For multiple choice items, use four or five choices whenever possible; make the distractors (incorrect responses) plausible and logical, but clearly incorrect (no trick questions).

g. In the writing of items, should a desired capability come to mind for which no topic objective was previously prepared, develop the necessary test items, and in the space for T.O. number write the word "New."

h. Proceed through the hierarchy of job skills and jobs in sequential order.

i. If assessment of goal attainment is by means of a checklist or rating scale, the test item form sheet will be used to establish the objective quantified measurement scale by which criterion performance will be evaluated.

6. Some principles that will help in the construction of items to measure proficiency.

a. The item as a whole should be realistic and practical; it should call for knowledge the student must use, or present a problem he may have to solve on the job.

b. The item should deal with an important and useful aspect of the job; leave out the trivia and useless information.

c. The item should be phrased in the working language; do not copy it from a manual or other test.

d. The item should be concerned with a capability required by the job.

e. Each item should be independent of other items; it should not be possible to answer an item based only on the content of some other item.

f. The item should be specific and deal directly with the job.

g. The central problem (item stem) should be clear and concise.

h. The problem should be stated accurately and precisely.

i. The problem should include all of the information needed but should be stated briefly.

j. The problem should contain only material relevant to its solution.

k. The distractors should be important, plausible answers; they should present common errors and misconceptions rather than trivial, illogical alternatives.

l. The best (correct answer) should not be given away by irrelevant details; the distractors should contain no extraneous clues.

m. The alternatives should deal with similar ideas or data expressed in parallel form.

SAMPLE FORM 6 (Proficiency Test Item)

VOC. AREA: Metals and Machines — TASK NO.: 9

FAMILY : Machines — C.O. NO.: 1

JOB : Surface Grinder — T.O. No.: 2

To grind a material of low tensile strength, it would be best to use the _____ abrasive type wheel.

A - Aluminum Carbide
B - Aluminum Oxide
C - Ferris Oxide
* D - Silicon Carbide

FORM 6 (Proficiency Measures) — Project ABLE

VOC. AREA: Metals and Machines — TASK NO.: 9

FAMILY : Machines — C.O. NO.: 1

JOB : Surface Grinder — T.O. No.: New

Identify the one set of metals that includes no metal of high tensile strength.

A - Lead, aluminum, tungsten
B - Aluminum, brass, steel
* C - Brass, cast iron, magnesium
D - Copper, molybdenum, bronze

APPENDIX B

DESCRIPTION OF TYPES OF QUESTIONS[1]

In the writing of test items, knowledge may be measured in several ways. Some knowledge can be approached in only one way, but as job behaviors are sampled at successively higher levels, it may be not only appropriate, but even necessary to write several items around a single capability (topic objective), each asking a different kind of question about that single behavior. The material below is intended to be descriptive and illustrative of the kinds of tasks that can be set by multiple-choice items.

Types of Items

1. Definition: What means the same as _____?
   Which of the following expresses the principle of _____ in different terms?

2. Purpose: What purpose is served by _____?
   What is the function of the _____?
   Why is this operation performed _____?
   What is the main reason for _____?
   Which of the following is an example of _____?

3. Cause: What is the cause _____?
   Under what condition is _____ true?

4. Effect: What is the effect of _____?
   If _____ is done, what will happen?

---

[1] Adapted from Mosier, C. I., Myers, M. C., & Price, Helen G. Suggestions for the Construction of Multiple-Choice Test Items. Educational & Psychological Measurement, Vol. 5, No. 3, Autumn 1945, as reported in Adkins, Dorothy C. Construction and analysis of achievement tests. Washington: U. S. Government Printing Office, 1947.

5. Association: What will occur at the time _____?

6. Recognition of Error: Which of the following represents an error in _____?

7. Identification of Error: What kind of error is _____?
   What name is given to the error in _____?
   What principle is violated when _____?

8. Evaluation: What is the best way to evaluate _____?
   For what reason is _____ the best evaluation?

9. Difference: What is the most important difference between _____?
   What feature best differentiates _____?

10. Similarity: What single characteristic makes for similarity between _____?

11. Arrangement: What is the proper order to meet the required sequence?
    Which of the following comes first in operating the _____?
    What is the next step after _____?

12. Incomplete Arrangement: What step has been omitted from _____?
    Where should step _____ be placed?

13. Common Principle: The following items except one are related by a common principle:
    What is the principle?
    Which _____ does not belong?
    Which of the _____ could be substituted to make all items related?

14. Controversial Subjects: Although there is not complete agreement on _____, what is the primary reason given by those who do support its desirability?